ARTICLE

DBKE-3NC: A DISTRIBUTED NEAREST-NEIGHBOUR NORMALIZATION ALGORITHM FOR DISTRIBUTED DATA MINING UNDER PRIVACY CONSTRAINTS

S. Urmela*, M. Nandhini

Department of Computer Science, Pondicherry University, INDIA

ISSN 0976-3104

ABSTRACT

Background: *To propose an architecture for DDM using Nearest Neighbour Normalization under privacy constraints for Electronic Health Records (EHRs) and Agriculture Weather Forecast.* ***Methods:*** *This paper proposed two algorithms are proposed: Nearest Neighbour Normalization with correlation (3NC) algorithm to normalize the raw data in local level (distributed datasite) and Double Blind Key-attribute based Encryption (DBKE) algorithm to deploy the protection of data in both local and global (central site) levels.* ***Results:*** *The proposed 3NC algorithm aims to maximize the accuracy and minimize the error rate by lossless and predicted normalization technique on understanding the data distribution in local levels. The proposed DBKE algorithm aims to maximize the confidential level and accuracy by efficient privacy technique. Further, proposed architecture aims to minimize the computational complexity and memory overhead by maintaining the history of records clusters and retrieving required clusters by proposed recall grading algorithm.* ***Conclusions:*** *Experimental implementation on EHRs and Agriculture Weather Forecast depicts an improved performance compared to other state-of-arts heterogeneous-distributed retrieval approaches.*

KEY WORDS
Distributed Data Mining, heterogeneous datasites, data normalization, privacy-preserving, Electronic Health Records, Agriculture Weather Forecast

INTRODUCTION

Reduction in storage cost and due to enormous growth of technologies has led to emergence of distributed systems. Most of the commercial off-the-shelf components are designed to work on centralized location. This centralized approach doesn't work well in recent distributed environments due to computation and storage constraints. Recent data mining involves database in which data is stored in one geographical location. Future data mining tasks works on data stored in different geographical location called Distributed Data Mining (DDM). The ultimate goal is to mine data which is distributed in homogeneous and heterogeneous sites [1].

Data located at local level is first mined with suitable DM algorithm and all the computed data from local level is agglomerated to form global level prediction. There are two-variations in this case, computation done at local level and data stored at global level [Fig. 1(a)], data stored at local level and computation done at global level [Fig. 1(b)]. A special variation of DDM, in which both data and computation is done at local level [2][Fig. 1(c)]. In proposed work, data and computation is stored both in local and global levels [Fig. 1(d)].

Alfredo Cuzzocrea (2013)[3] stated that framing a framework/methodology for DDM is challenging not only by distributed environment, but also for its minimized computational complexity and efficient storage of data in local and global levels. Centralized data mining algorithms/systems designed for centralized systems cant' be applied for distributed environment. Fu Y et al. (2012)[4] discussed certain key characteristics in designing DDM algorithms: minimizing space and computation complexity, designing suitable algorithm for both homogeneous and heterogeneous datasets and maintaining local datasets independence. Further, all these issues are interrelated to each other. This has set way to many researchers to carry-out their work in this field. The main contributions of proposed work are summarized as follows:

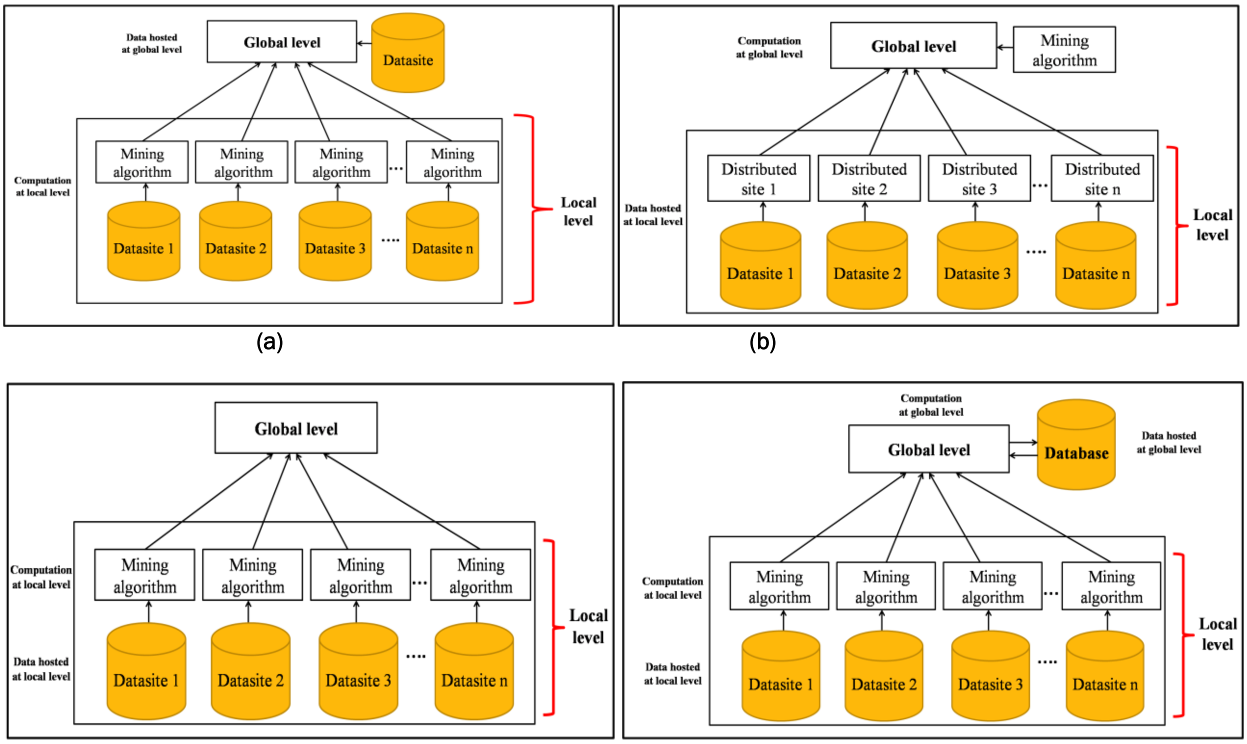
Received: 18 Sep 2018
Accepted: 30 Nov 2018
Published: 02 Jan 2019

- Proposing an effective privacy-preserving distributed mining algorithm based on Double-Blind Key Attribute based Encryption algorithm with Nearest Neighbour Normalization-Correlation which successfully utilizes the distributed environment.
- Proposing an effective heterogeneous distributed mining technique which maintains universal dataset across distributed datasites.
- Design and evaluation of proposed mining technique on Electronic Health Records (EHRs) and Agriculture Weather Forecast dataset for resolving data disclosure and for predicting dynamic updations in distributed environment.
- Measuring the metrics such as effectiveness, efficiency, privacy-technique evaluation and normalization technique evaluation metrics on proposed technique.

The proposed architecture in this paper involves records retrieved by Nearest Neighbour Normalization-Correlation (3NC) technique with Double-Blind Key Attribute based Encryption algorithm (DBKE). At local level, distributed data stored in multi-linked list is privacy preserved by DBKE algorithm. DBKE algorithm prevents key-attribute data disclosure to neigbouring local levels and global level by encrypting the data with double-blinded function (symmetric encryption function and hashing function). At global level, encrypted key-attribute in decrypted by reverse encryption (unhashing function and symmetric decryption function). This privacy algorithm is for preventing data disclosure during storage in local levels. The key-attribute is used for classifying records. 3NC technique uses MAX and MIN value of key-attribute for

***Corresponding Author**
Email:
urmelaindra@gmail.com

forming dendrogram. Say, for classifying a patient diagnosed with corresponding disease or not. Initially, the prime key-attribute is used for classifying on analyzing the MAX and MIN value. From the clusters formed, MAX value is normalized to highest value in corresponding cluster. Then by correlation, secondary key-attribute is used for further classifying patients diagnosed with particular disease or not. From the sub-clusters formed, MAX value is normalized to highest value in corresponding sub-cluster. Each branch of dendrogram corresponds to a cluster (sub-cluster). The clusters formed are migrated to global level for final result prediction.



(a) (b)

(c) (d)

Fig. 1: Working architecture – DDM (a) data hosted at global level, computation at local level (b) data hosted at local level, computation at global level (c) both data hosting and computation in local level (d) data hosting and computation in both local and global levels.

At global level, old clusters migrated from local levels assigned grades by proposed recall grading algorithm. The number of clusters formulated for each disease in local levels (on-the-fly) and stored in global level (old records) must be equal. Clusters communicated to global level were assigned grades by recall grading algorithm. Finally, both cluster sets retrieved from local levels and records stored in global level (old records) sums up their grades and their average value is again assigned final grades. Proposed architecture is implemented on real-world openly available datasets taken from UCI data repository.

The organization of paper is as follows: Section 2 discusses EHRs and Agriculture weather forecast dataset description. Section 3 depicts proposed architecture for effective retrieval of records in distributed environment. Section 4 presents performance analysis and result discussion of proposed architecture. Section 5 summarizes the paper.

DATASET DESCRIPTION

EHRs and Agriculture Weather Forecast are the datasets considered for evaluation of proposed architecture. Dataset description of both the datasets is discussed here.

Case study 1: Electronic Health Records (EHRs)

EHRs are becoming worldwide popular way of maintaining health records of a person. The difference between EHR and Patient Health Record(PHR) is that PHR includes medical record for only a particular disease diagnosed for a person (for a short time period say, 3 years) whereas EHR includes medical record of a person right from their birth till death (for lifetime)[5]. EHR includes person therapeutic history, drug usage, allergies and test outcomes. It is supported by EHR management and decision support [6]. In DDM, EHRs are utilized and updated in either homogeneously/heterogeneously distributed data sites.

Case study 2: Agriculture Weather-Forecast

Agriculture database uses information collected from different geographical location in periodical manner. Details relating to crop plantation and weather monitored data collected from heterogeneous data sites are maintained in a universal database. This profile includes plantation details, weather details relating to heterogeneous sites. This data is used for giving efficient and timely farming details to farmers. [7]

Analysis of case studies

Both the case studies considered are analyzed on different parameters as shown in [Table 1] stakeholders, expert systems used, local and global level DDM users, no. of local levels considered for evaluation of proposed algorithm, profiles of case study with commonality, common attributes and key attributes of applications used for formulating dendrogram.

Table 1: Comparison of case study 1: EHRs and case study 2: Agriculture weather forecast

Parameters	Electronic Health Records (EHRs)	Agriculture Weather Forecast
Stakeholders	Health care providers (doctors, nurses, etc)	Weather forecaster, Agriculture specialist
	Patients	Agriculturist
	Health experts	Weather and Agriculture experts
Expert systems	EHR management and decision support	Weather Forecast record management and decision support
DDM-Local level users	Health care personnel	Agriculturist personnel
DDM-Global level users	Health care personnel/other government personnel related to Health	Weather forecaster and agriculture specialist Personnel
No. of local levels	7	7
Profiles	Demographic profile	Demographic profile
	Anthropometric profile	Soil profile
	Clinical results profile	Weather profile
	Medication/allergies vaccination profile	Plantation profile
Common attributes	Patient Name Age Address Gender Location	Farmer Name Age Address Gender Location
Key-attributes	Hypertension: SBP DBP Age Diabetes: GlyHb (Glycalated Hemoglobin) Age	Agriculture Weather Forecast: Relative_temperature Humidity Wind Speed Age
Functionality	Adding new EHR record at local level	Adding new farmer record at local level
	Updating existing EHR record at local level	Updating existing farmer record at local level
	Deleting EHR record at local level	Deleting farmer record at local level
	Filtering EHR record based on user query at global level	Filtering farmer record based on user query at global level
	Dendrogram formulation with EHRs at local level based on user query	Dendrogram formulation with farmer records at local level based on user query

DESIGN OF PROPOSED ARCHITECTURE

Proposed architecture working is classified at two-levels: local and global levels. At local level records undergo proposed two stages namely,

STAGE I: Data Mining (Nearest Neighbour Normalization-Correlation algorithm)
STAGE II: Privacy-preserving of local and global data (Double-Blind Key-attribute based Encryption algorithm)

At global level records undergo proposed two stages namely,

STAGE I: Recall-grading algorithm
STAGE II: Filtering final cluster set

STAGE I: Data Mining (Nearest Neighbour Normalization-Correlation algorithm)

Filling missing values

In proposed architecture, at local levels the data structure used for storage of data is multi-linked list. Consider EHRs dataset, the data storage at local level is depicted in fig. 2. At each local level, the missing values are filled from the relative data of corresponding entity. For any type application, there are two types of filling missing values. Filling demographic data and filling application-specific data[8]. Demographic data includes filling general information. Example: filling entity age from DOB, etc. Application-specific data includes filling application-oriented data8. For medical dataset, clinical results of patient are filled by relative computation of corresponding data. For example, say in hypertension dataset if the value of SBP (Systolic Blood Pressure)/DBP (Diastolic Blood Pressure) for a continuous range is 148/78, 140/79, 152/73 then the next value is computed as relative mean (147/77)[9].

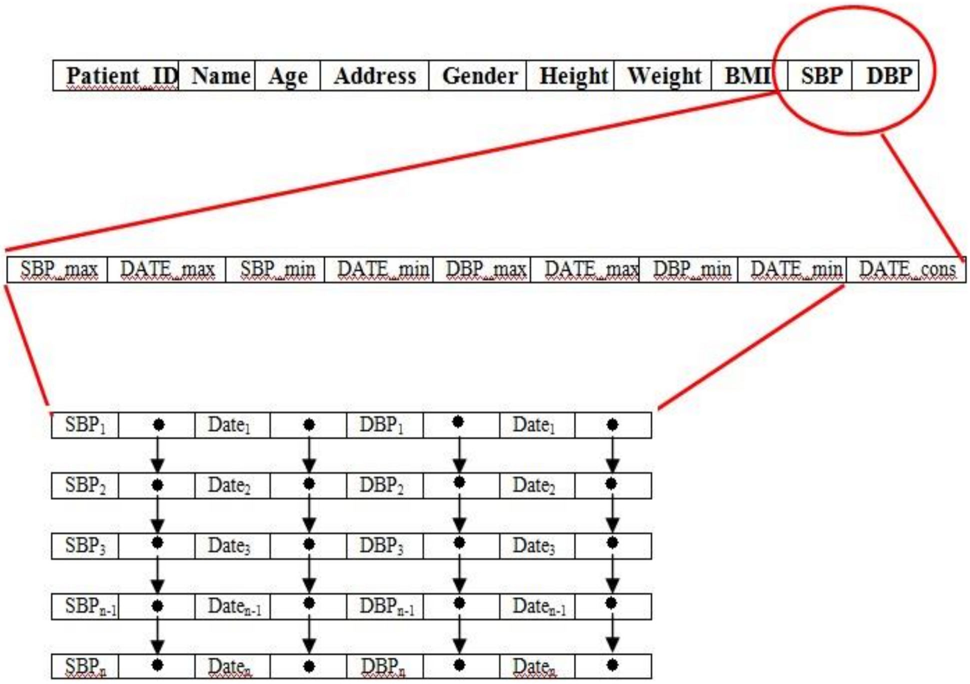


Fig. 2: Data Structure representation

……………………………………………………………………………………

Say, for a patient at each local level, if SBP/DBP value is monitored daily, it leads to huge volume of data values. To overcome the problem of maintaining history of records, for every 20 data values of SBP and DBP database is consolidated. MAX and MIN value of SBP and DBP among the old 20 records along with time recorded of MAX and MIN values and the time data was consolidated is noted. In next sub-stage, key attributes (SBP and DBP) is normalized. Key-attributes are sensitive attributes used for classifying patients diagnosed with disease or not which needs to be privacy preserved. In this proposed architecture, SBP and DBP values are used for classifying a patient as hypertension or not.

Dendrogram formulation by Nearest Neighbour Normalization-Correlation (3NC)

From the clinical data, MAX and MIN values of key-attributes are calculated using which clusters are formulated for mining. For example if SBP value for a single entity reading is 120,134,115,132,134,121,117,128 then MAX = 134 and MIN = 115. Similarly DBP value reading is 80,79,83,82,82,84,82,80 then MAX = 84 and MIN = 79. ICD-9 code (International statistical code for Classification of Diseases)[10] is used as a standard reference for classifying patients diagnosed with disease.

Consider the case of classifying patient diagnosed with hypertension disease. There are two-levels involved in formulating dendrogram. At first level, the dendrogram is formulated with SBP values as shown in fig. 3. According to ICD-9 code, the range of classifying SBP is <=120, 121-140,141-160,161-180,181-200 and >200 on analyzing the (MAX,MIN) pair. From the fig. 3 on first level classification clusters are formed say C1 to C24 with defined range. In C1, patients (P1-P10) fall under category with SBP <=120. From the (MAX,MIN) pair of 10 patients, all 10 MAX value is normalized to largest MAX value in corresponding cluster. Likewise for all 24 clusters, same normalization technique is followed.

At second level, from the normalized clusters second level of classifying patients with DBP is done. According to ICD-9 code, the range of classifying DBP is <=80, 81-100,101-120, >120 on analyzing the (MAX,MIN) pair. The way of classifying records in two-way as discussed above is called correlation. Correlation among two variables means they are interdependent on each other. For diagnosing a patient with hypertension, both SBP and DBP values are monitored. Hence the algorithm proposed is called 3NC

IIOAB JOURNAL
ISSN: 0976-3104

algorithm. From the [Fig. 3] on second level classification sub-clusters with defined range of DBP is formed. In C11, patients (P2, P5 and P6) fall under category with DBP <=80. From the (MAX,MIN) pair of those three patients, all three MAX value is normalized to largest MAX value of DBP in corresponding cluster. Likewise for all sub-clusters, same normalization technique is followed. According to the example in [Fig. 3], 96 sub-clusters are formed by 3NC which in-depth classifies patients with hypertension disease.

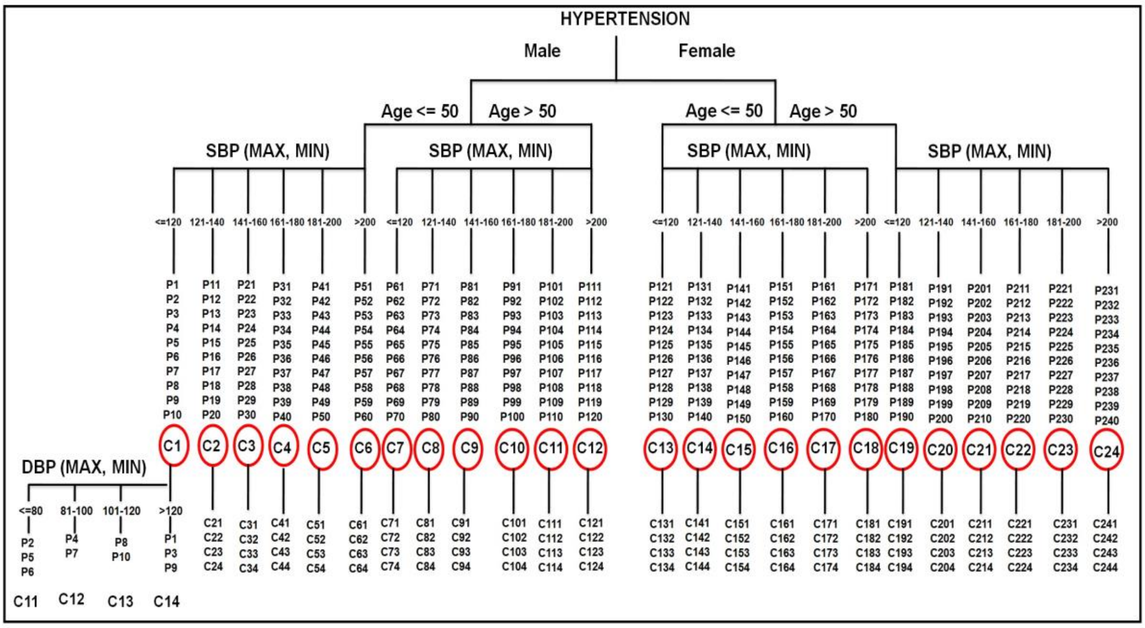

HYPERTENSION
Male
Female
Age <= 50
Age > 50
Age <= 50
Age > 50
SBP (MAX, MIN)
SBP (MAX, MIN)
SBP (MAX, MIN)
SBP (MAX, MIN)
<=120 121-140 141-160 161-180 181-200 >200
<=120 121-140 141-160 161-180 181-200 >200
<=120 121-140 141-160 161-180 181-200 >200
<=120 121-140 141-160 161-180 181-200 >200
P1 P2 P3 P4 P5 P6 P7 P8 P9 P10
P11 P12 P13 P14 P15 P16 P17 P18 P19 P20
P21 P22 P23 P24 P25 P26 P27 P28 P29 P30
P31 P32 P33 P34 P35 P36 P37 P38 P39 P40
P41 P42 P43 P44 P45 P46 P47 P48 P49 P50
P51 P52 P53 P54 P55 P56 P57 P58 P59 P60
P61 P62 P63 P64 P65 P66 P67 P68 P69 P70
P71 P72 P73 P74 P75 P76 P77 P78 P79 P80
P81 P82 P83 P84 P85 P86 P87 P88 P89 P90
P91 P92 P93 P94 P95 P96 P97 P98 P99 P100
P101 P102 P103 P104 P105 P106 P107 P108 P109 P110
P111 P112 P113 P114 P115 P116 P117 P118 P119 P120
P121 P122 P123 P124 P125 P126 P127 P128 P129 P130
P131 P132 P133 P134 P135 P136 P137 P138 P139 P140
P141 P142 P143 P144 P145 P146 P147 P148 P149 P150
P151 P152 P153 P154 P155 P156 P157 P158 P159 P160
P161 P162 P163 P164 P165 P166 P167 P168 P169 P170
P171 P172 P173 P174 P175 P176 P177 P178 P179 P180
P181 P182 P183 P184 P185 P186 P187 P188 P189 P190
P191 P192 P193 P194 P195 P196 P197 P198 P199 P200
P201 P202 P203 P204 P205 P206 P207 P208 P209 P210
P211 P212 P213 P214 P215 P216 P217 P218 P219 P220
P221 P222 P223 P224 P225 P226 P227 P228 P229 P230
P231 P232 P233 P234 P235 P236 P237 P238 P239 P240
C1 C2 C3 C4 C5 C6 C7 C8 C9 C10 C11 C12
C13 C14 C15 C16 C17 C18 C19 C20 C21 C22 C23 C24
DBP (MAX, MIN)
<=80 81-100 101-120 >120
P2 P5 P6
P4 P7
P8 P10
P1 P3 P9
C11 C12 C13 C14
C21 C22 C23 C24
C31 C32 C33 C34
C41 C42 C43 C44
C51 C52 C53 C54
C61 C62 C63 C64
C71 C72 C73 C74
C81 C82 C83 C84
C91 C92 C93 C94
C101 C102 C103 C104
C111 C112 C113 C114
C121 C122 C123 C124
C131 C132 C133 C134
C141 C142 C143 C144
C151 C152 C153 C154
C161 C162 C163 C164
C171 C172 C173 C174
C181 C182 C183 C184
C191 C192 C193 C194
C201 C202 C203 C204
C211 C212 C213 C214
C221 C222 C223 C224
C231 C232 C233 C234
C241 C242 C243 C244

Fig. 3: Dendrogram formulation - Nearest Neighbour Normalization-Correlation

……

[Fig. 3] depicts correlation and normalization on EHRs for a single cluster C1 retrieved from dendrogram. Initially cluster set is retrieved on analyzing the (MAX,MIN) pair of SBP value. After forming cluster with SBP value, MAX value is normalized for all the clusters formed. Then DBP value is analyzed which forms sub-cluster with (MAX,MIN) pair of DBP value. Consider for C1, four sub-clusters are formed. Sub-clusters formed for C1 are C11,C12,C13,C14. After sub-clusters formation with SBP and DBP, MAX value of DBP in each cluster is normalized with highest MAX value of corresponding cluster set.

STAGE II: Privacy-preserving of local data (Double-Blind Key-attribute based Encryption algorithm)

The calculated (MAX,MIN) pair is highly-sensitive data which needs to be protected and kept confidential at local level and during migration of data from local level to global level for resultant query. The value of (MAX,MIN) pair of key attribute is used for mining. The proposed Double-Blind Key-attribute based Encryption (DBKE) algorithm at each local level employs an encryption function followed by hashing function. The key used for encryption function and hashing function of each local level is stored as a secret share between each heterogeneous distributed local and global level.

Blind function is a service provided to client which computes function for encrypting the data without revealing the original data [11]. In proposed algorithm, Double-Blind technique is followed which computes two encryption functions at local level and two decryption functions at global level to protect sensitive key-attribute data.

Processing at local level

The (MAX,MIN) pair computed data of key-attribute is twice privacy-preserved with an encryption function and a hashing function.

Encryption function by symmetric encryption key

At first level, the (MAX,MIN) value of key-attribute is encrypted using symmetric encryption key.
The encryption function is computed as,
Encrypted data, E1 = E (K,I)

Encrypted data, E1 = E (K,(MAX,MIN))
where I - input data (MAX,MIN) and K - key
value for encryption.
Hashing function by key-value

At second level, hashing function is applied to the encrypted data, which computes hashing function individually for MAX and MIN of key-attribute.
Hash function, F(E1) = E1 MOD N
where E1 – Encrypted data and N – key value for encryption.

Processing at global level

The Double-Blind encrypted (MAX,MIN) pair migrated from local level is decrypted in reverse. At first level, unhashing function (Reverse decryption) is applied on double-blind encrypted (MAX,MIN) data.

Unhashing function by reverse decryption

On knowing the N, key value used in hashing function, reverse decryption process is possible which computes the MAX and MIN value individually. The output of unhashing function is D1.

Decryption function by symmetric decryption key

At second level, decryption function is applied on the single-decrypted (MAX,MIN) data to obtain the original (MAX,MIN) pair.
The decryption function,
Input data, I = D (D1,K)
where, K – key value used for encryption at local level, D1 – single-level decryption output (Unhashing function).

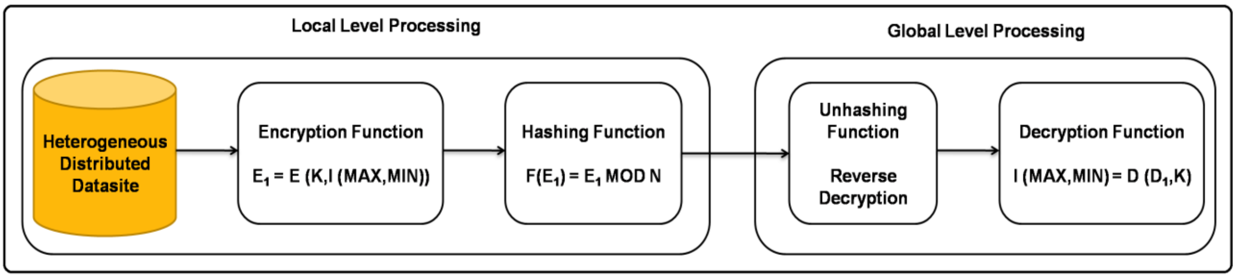


Fig. 4: Double-Blind Key-attribute based Encryption algorithm

STAGE I: Recall-Grading Algorithm

At global level, record clusters retrieved by 3NC algorithm undergoes proposed recall-grading algorithm. Some clusters (old records) would be retrieved earlier by 3NC algorithm. With those clusters a matrix is formulated with clusters say, C11, C21, etc and user-query Q1, Q2, etc. A ✓(tick mark) indicates including the corresponding clusters in user-query and X (cross mark) indicates excluding the corresponding clusters in user-query. Only those records which are updated in local level or new records corresponding to user query, undergoes 3NC algorithm and clusters are transferred from local level to global level. Based on proposed recall-grading algorithm, grades are assigned for each corresponding clusters (from same dendrogram branch) based on recall metric.

Recall-grading algorithm for each cluster takes on by calculating recall metric of each records of individual clusters. The calculated recall metric of each cluster is obtained by taking an average of summing individual records recall metric. Grades are assigned to calculated recall metric. Higher recall value were assigned higher grades.
Each record cluster transferred from local levels to global level is considered. Recall metrics of individual records of each cluster is calculated. Recall metric of cluster is summation of recall metric of individual records and taking an average. Finally, grades are assigned to individual cluster. Higher grades being assigned to higher recall metric clusters. Higher grades cluster corresponds to relevant records corresponding to user query Q.

STAGE II: Filtering final cluster set

Record cluster sets of old records and new records along with their grades are finally consolidated by taking an average of both grades. Architecture of proposed DBKE-3NC, distributed Nearest-Neighbour Normalization algorithm for DDM under privacy constraints is shown in [Fig. 5].

EXPERIMENTAL IMPLEMENTATION

The experiment is carried out on a single 64-bit machine, windows 7 OS having 3GHz Intel dual core processor with 4GB main memory. The proposed algorithm is coded in C# and implemented in HDFS

(Hadoop Distributed File System) distributed environment. The proposed architecture is implemented with seven distributed local levels both for EHRs and Agriculture Weather Forecast dataset. Seven local levels are created with a server at global level. Comparative analyses have been carried out with performance evaluation metrics of proposed architecture with conventional approaches. The proposed architecture is implemented on openly-available real-world EHRs and Agriculture Weather Forecast obtained from UCI repository [12].

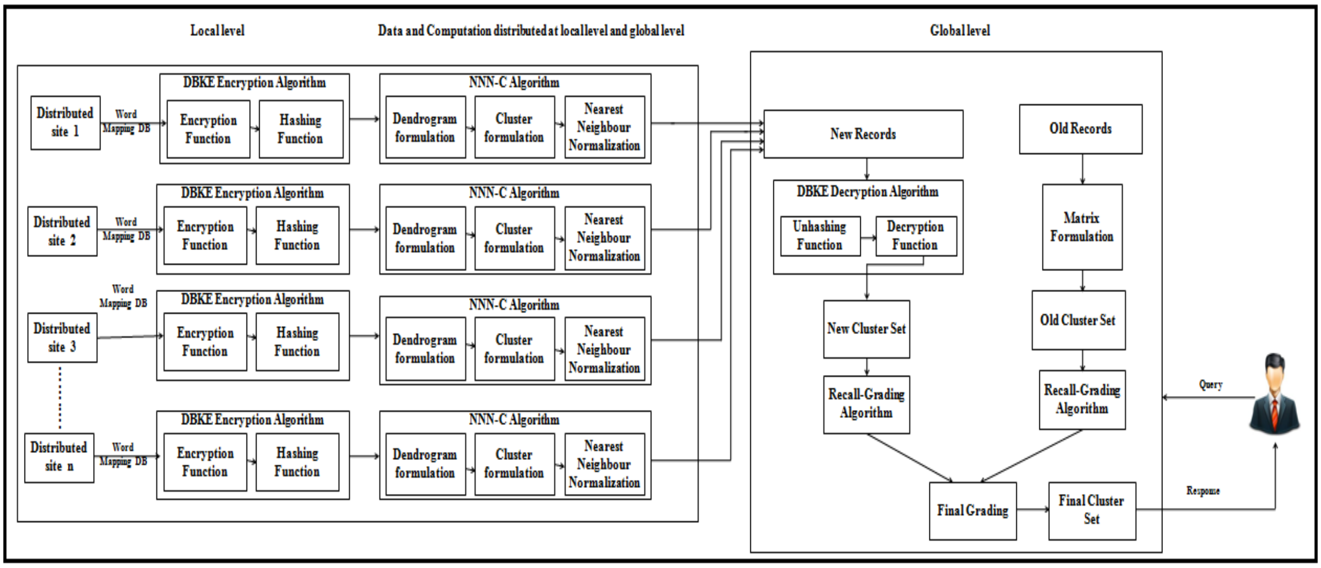


Fig. 5: Proposed Architecture - DBKE-3NC: distributed Nearest-Neighbour Normalization algorithm for Distributed Data Mining under privacy constraints.

....

Effectiveness metrics

This measure is computed in terms of retrieving successful result meeting user query. Effectiveness measures include precision[13], recall[14], F-measure[15] and result accuracy[16].

Table 2: Effectiveness measures Vs Heterogeneous classifier approaches

Dataset	Metrics (%)	Collective Decision Tree	Distributed Clustering	Collective Bayesian Learning	Collective Principal Component Analysis	Proposed DendrogramCluster formulation
EHRs	Precision	0.6942	0.9218	0.9012	0.9045	0.9434
	Recall	0.6571	0.9202	0.8449	0.8756	0.9023
	F-measure	0.6751	0.9103	0.8721	0.8937	0.9224
Agriculture Weather Forecast	Precision	0.6582	0.9190	0.8390	0.8759	0.9381
	Recall	0.6337	0.8681	0.8054	0.8571	0.9010
	F-measure	0.6482	0.8976	0.8189	0.8668	0.9218

[Table 2] depicts precision, recall and F-measure comparison of proposed mining algorithm with the state-of-art mining techniques. Collective Principal Component Analysis (P-90.45%, R-87.56% for EHRs and P-87.59% R-85.71% for Agriculture Weather Forecast) exhibits more compared to former two techniques because dynamic updations of records from individual local levels are considered but lesser compared to proposed dendrogram cluster formulation technique (P-94.34% R-90.23% for EHRs and P-93.81% R-90.1% for Agriculture Weather Forecast) because repetitive records are considered which needs to be eliminated leading to decreased retrieval rate. Collective Bayesian learning exhibits more values (P-90.12% R-84.49% for EHRs and 83.9% R-80.54% for Agriculture Weather Forecast) compared to Collective decision tree exhibits (P-69.42% R-65.71% for EHRs and 65.82% R-63.37% for Agriculture Weather Forecast) because in bayesian learning technique on-the-fly or updated records are considered for mining whereas in distributed clustering updated records are considered. Distributed clustering exhibits higher values (P-92.18% R-90.02% for EHRs and P-91.9% R-86.81% for Agriculture Weather Forecast) compared to all the techniques because dynamic records are considered for mining leading to effective retrieval rate but less than proposed because of mismatch of key-attributes of each local level.

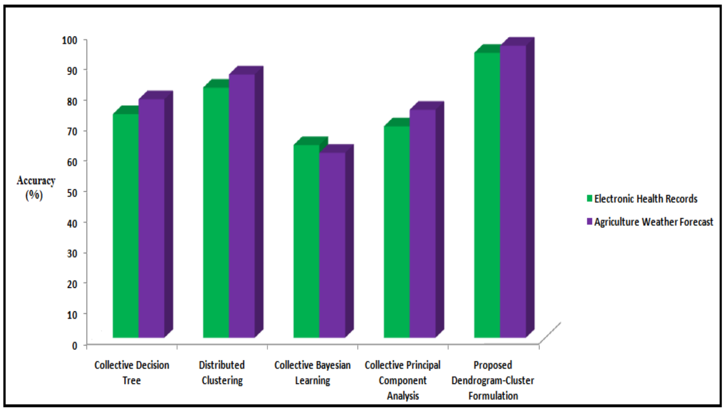


Fig. 6: Accuracy Vs Heterogeneous classifier approaches.

[Fig. 6] depicts accuracy comparison of proposed mining algorithm with the state-of-art mining techniques. Proposed dendrogram cluster formulation technique exhibits more accuracy (93.56% for EHRs and 95.89% for Agriculture Weather Forecast) compared to former four techniques because updated records from local levels is considered and redundant records are cross-checked. Distributed clustering (82.14% for EHRs and 86.47% for Agriculture Weather Forecast) exhibits more accuracy compared to collective Principal Component Analysis (69.41% for EHRs and 74.87% for Agriculture Weather Forecast), collective Bayesian learning (63.28% for EHRs and 60.72% for Agriculture Weather Forecast) and collective decision tree (73.4% for EHRs and 78.29% for Agriculture Weather Forecast) because of two-level of dendrogram formulation leading to sub-cluster formulation which increases the result accuracy

Efficiency metrics

This measure is computed in terms of retrieving effective result on varying environment. Efficiency measures include execution time [17] and memory cost [18].

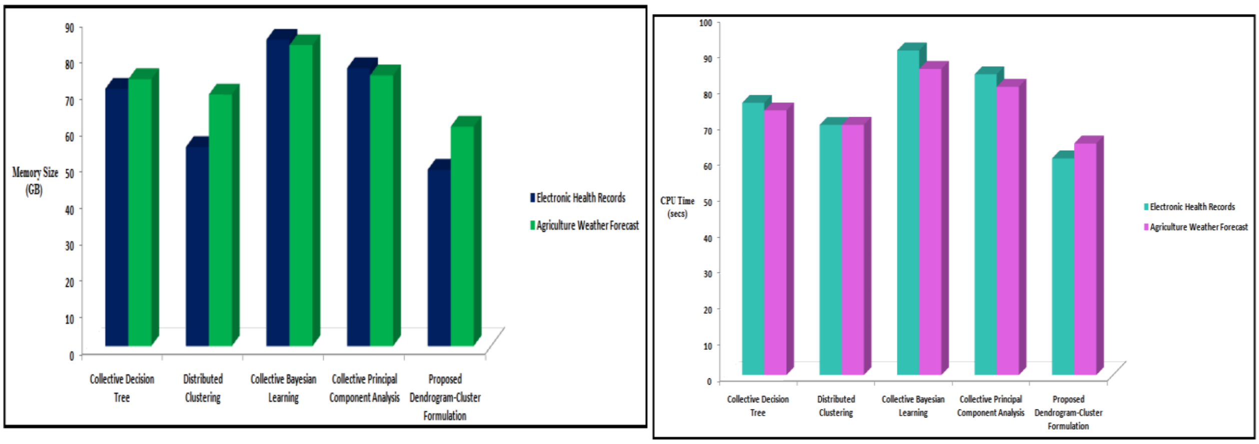


Fig. 7: (a) Memory Size Vs Heterogeneous classifier approaches (b) Execution Time Vs Heterogeneous classifier approaches.

[Fig. 7] depicts memory size comparison of proposed mining algorithm with the state-of-art mining techniques. Proposed dendrogram cluster formulation technique exhibits less memory cost (48.54% for EHRs and 60.39% for Agriculture Weather Forecast) compared to former four techniques because old key-attribute values are consolidated for every 20 values and by normalization technique data values are minimized leading to less memory cost. Distributed clustering exhibits less memory size (54.78% for EHRs and 69.25% for Agriculture Weather Forecast) compared to collective Principal Component Analysis (76.48% for EHRs and 74.52% for Agriculture Weather Forecast) collective Bayesian learning (84.34% for EHRs and 82.83% for Agriculture Weather Forecast) collective decision tree (70.81% for EHRs and 73.48% for Agriculture Weather Forecast) because by clustering technique old records are stored in global level and only new records are migrated from local level to global level for

[Fig. 7] depicts execution time comparison of proposed mining algorithm with the state-of-art mining techniques. Proposed dendrogram cluster formulation technique exhibits less execution time (60.43% for EHRs and 64.6% for Agriculture Weather Forecast) compared to former four techniques because by two-level dendrogram and sub-cluster formation time taken to retrieve records corresponding to user query is reduced. Further, by word-mapping database at global level for identifying the key-attribute to be considered at each local level for mining reduces the time taken for mining process. Distributed clustering (69.78% for EHRs and 69.81% for Agriculture Weather Forecast) exhibits less execution time compared to other three techniques, (collective Principal Component Analysis (83.89% for EHRs and 80.41% for Agriculture Weather Forecast) collective Bayesian learning (90.54% for EHRs and 85.39% for Agriculture Weather Forecast) collective decision tree (75.92% for EHRs and 73.81% for Agriculture Weather Forecast)) because clustering technique retrieves old key-attribute entries and new key-attribute entries are retrieved for every updations.

Privacy technique evaluation metrics

Performance of proposed privacy-preserving algorithm is analyzed on two measures: accuracy and confidence level [19].

Table 3: Privacy evaluation metrics Vs Privacy techniques

Dataset	Metrics (%)	Additive Noise Addition	Multiplicative Noise Addition	L-Diversity	T-Closeness	DBKE Algorithm
EHRs	Accuracy	98.23	99.32	97.14	96.73	99.38
	Confidence Level	94.37	96.21	98.37	99.42	99.40
Agriculture Weather Forecast	Accuracy	97.59	98.37	96.32	97.02	99.04
	Confidence level	94.89	95.92	97.58	98.74	98.91

[Table 3] depicts accuracy comparison of proposed DBKE algorithm with the state-of-art privacy preserving approaches. Proposed DBKE algorithm exhibits more accuracy (99.38% for EHRs and 99.04% for Agriculture Weather Forecast) compared to other randomization and anonymization approaches because of combining both randomization and anonymization approach. Randomization approaches, additive noise addition (98.23% for EHRs and for 97.59% Agriculture Weather Forecast) and multiplicative noise addition (99.32% for EHRs and 98.37% for Agriculture Weather Forecast) exhibits more accuracy value but privacy/confidence level is less because original data can be retrieved on analyzing the spectral values of randomized data. Anonymization approaches, L-Diversity (97.14% for EHRs and 96.32% for Agriculture Weather Forecast) and T-Closeness (96.73% for EHRs and 97.02% for Agriculture Weather Forecast) exhibits less accuracy compared to randomization approaches but confidence/privacy level is more than randomization approaches. In conventional approaches there is some data loss incurred while converting the original data to protected data.

[Table 3] depicts confidence level comparison of proposed DBKE algorithm with the state-of-art privacy preserving approaches. Proposed algorithm exhibits more confidence level (99.40% for EHRs and 98.91% for Agriculture Weather Forecast) compared to randomization and anonymization approaches because the distribution of original data is considered in proposed DBKE algorithm. Randomization approaches, additive noise addition (94.37% for EHRs and 94.89% for Agriculture Weather Forecast) and multiplicative noise addition (96.21% for EHRs and 95.92% for Agriculture Weather Forecast) exhibits less confidence level compared to anonymization approaches and proposed DBKE algorithm. Similarly, anonymization approaches, L-Diversity (98.37% for EHRs and 97.58% for Agriculture Weather Forecast) and T-Closeness (99.42% for EHRs and 98.74% for Agriculture Weather Forecast) exhibits higher or equal confidence level compared to proposed DBKE algorithm but result accuracy is less. In conventional approaches data distribution at each local datasite and global level is not considered while converting the original data to protected data.

Normalization technique evaluation metrics

Performance of proposed normalization algorithm is analyzed on two measures: MSE and RMSE [20].

Table 4: Normalization evaluation metrics Vs Normalization techniques

Dataset	Metrics (%)	Min-max normalized data	z-score normalized data	Decimal scaling normalized data	3N without C normalized data	3NC normalized data
EHRs	MSE	1.2313	1.3621	1.1182	0.9045	0.1972
	RMSE	1.1782	1.1892	1.2781	0.9510	0.3958
Agriculture Weather Forecast	MSE	1.4431	1.5467	1.1576	1.7345	0.2056
	RMSE	1.2012	1.2437	1.0760	1.3171	0.4534

Table 4 depicts MSE and RMSE comparison of proposed 3NC algorithm with the state-of-art normalization techniques. Min-max exhibits less MSE and RMSE (1.2313, 1.1782 for EHRs and 1.4431, 1.2012 for Agriculture Weather Forecast) compared to z-score (1.3621, 1.1892 for EHRs and 1.5467, 1.2437 for Agriculture Weather Forecast) because out-of-bound error is triggered in min-max normalization when the normalized range doesn't fall within the specified range. In decimal scaling MSE and RMSE (1.1182, 1.2781 for EHRs and 1.1576, 1.0760 for Agriculture Weather Forecast) error rate is less compared to former two normalization techniques because prior knowledge of max and min value of normalized attribute is known. 3N without C algorithm (0.9045, 0.9510 for EHRs and 1.7345, 1.3171 for Agriculture Weather Forecast) exhibits more error rate compared to proposed 3NC because data distribution is not considered leading to less accuracy rate. Proposed 3NC algorithm (0.1972, 0.3958 for EHRs and 0.2056, 0.4534 for Agriculture Weather Forecast) exhibits better performance than former three state-of-art normalization techniques because data distribution at local levels is considered minimizing the error rate. Proposed 3NC algorithm considers all the data within specified range and outlier values for normalization.

CONCLUSION

In recent years, DDM evolved in large aiming for efficient and timeliness data retrieval. Many proposed works framed model for partial DDM, rather utilizing distributed data and computation. In this paper, a distributed Nearest-Neighbour Normalization algorithm for DDM under privacy constraints is proposed with dynamic distributed datasets, EHRs and Agriculture Weather Forecast. In this model, algorithms for normalization and privacy are proposed. By normalization-correlation technique, two-level dendrograms and sub-clusters are formed which leads to efficient classification of records under privacy constraints thereby achieving less memory cost and increased precision and recall values.

CONFLICT OF INTEREST
None

ACKNOWLEDGEMENTS
None

FINANCIAL DISCLOSURE
None

REFERENCES

[1] Das K, Bhaduri K, Gupta H K. [2017] A local asynchronous distributed privacy preserving feature selection algorithm for large peer-to peer networks. Journal of Knowledge and Information Systems, 24(3): 341-367.

[2] Urmela S, Nandhini M. [2017] Approaches and Techniques of Distributed Data Mining. International Journal of Engineering and Technology (IJET), 9(1): 63-76.

[3] Alfredo Cuzzocrea. [2017] Models and algorithms for high-performance distributed data mining. Elsevier Journal of Parallel and Distributed computing, 73(93): 281-283.

[4] Fu Y. [2012] Distributed Data Mining: An Overview. In: Newsletter of the IEEE Technical Committee on Distributed Processing, 5-9.

[5] Breiman L. [1996] Pasting small votes for classification in large databases and on-line. Machine Learning, 36:85-103.

[6] Atisha Sachan. [2016] A Survey on Recommender Systems based on Collaborative Filtering Technique. International Journal of Innovations in Engineering and Technology (IJIET), 2 (2): 8-14.

[7] Ganga Devi SVS. [2014] A Survey on Distributed Data Mining and Its Trends. International Journal of Research in Engineering & Technology (IJRET), 107-120.

[8] Josenildo Costa da Silva, Matthias Klusch. [2006] Inferences in Distributed Data Mining. Engineering Applications of Artificial Intelligence, 19:363-369.

[9] Kargupta, Kamath Chan. [1999] Distributed and Parallel Data Mining: Emergence, Growth and Future Directions. Advances in Distributed Data Mining, (eds.). Hillol Kargupta and Philip Chan AAAI Press, 407-416.

[10] Kawuu W Lin, Sheng-Hao Chung. [2006] A fast and resource efficient mining algorithm for discovering Frequent patterns in distributed computing environments. Journal of Future Generation Computer Systems, 52:49-58.

[11] Koren Y. [2015] Tutorial on recent progress in collaborative filtering. In Proc. of the 2nd ACM Conference on Recommender Systems, 8-67.

[12] [dataset] Job Recruitment Dataset. https://github.com/datameet.

[13] Park BH, Kargupta H. [2002] Distributed Data Mining: Algorithms, Systems, and Applications. In. Data mining handbook.

[14] Nandhini M, Urmela S. [2016] Clustered Collaborative Filtering Approach for Distributed Data Mining on Electronic Health Records. International Journal of Control Theory and Applications (IJCTA), 9(3):81-91.

[15] Baik S, Bala J. [2004] A Decision Tree Algorithm for Distributed Data Mining: Towards Network Intrusion Detection. Computational Science and Its Applications – (ICCSA), 3046:206-212.

[16] Pandey TK, Panda N, Sahu PK. [2012] Improving performance of distributed data mining (DDM) with multi-agent system. International Journal of Computer Science Issues (IJCSI), 9(2):74-82.

[17] Tsoumakas G, Vlahavas I. [2008] Distributed Data Mining. Encyclopedia of Data Warehousing and Mining, 709-715.

[18] Vinaya Sawant, Ketan Shah. [2013] A review of Distributed Data Mining using agents. International Journal of Advanced Technology & Engineering Research (IJATER), 3(5):27-33.

[19] Xiaoyuan Su, Khoshgoftaar Taghi M [2009] A Survey of Collaborative Filtering Techniques. Advances in Artificial Intelligence.

[20] Yan Li, Changxin Bai, Chandan K Reddy. [2016] A distributed ensemble approach for mining health care data under privacy constraints. Journal of Information Sciences, 330: 245-259.